ISSN 1682-296X (Print)
ISSN 1682-2978 (Online)

Bio Technology

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Biotechnology 8 (3): 329-342, 2009
ISSN 1682-296X

Functional Prediction of *Calamus manan* Inflorescence ESTs Through Motif Detection

[1]K. Nadarajah, [2]C.Y. Choong, [2]S.J. Leong and [2]R. Wickneswari
[1]School of Biosciences and Biotechnology,
[2]School of Natural and Environmental Sciences, Faculty of Science and Technology,
Universiti Kebangsaan Malaysia, 43600 UKM Bangi Selangor Darul Ehsan, Malaysia

Abstract: *Calamus manan* floral cDNA libraries were constructed for four stages of flowering in male and female plants, respectively. The *Calamus manan* inflorescence ESTs were generated to provided a better understanding of the flowering process through the identification of genes that are expressed in the floral tissues of this plant. The BLASTX homology search showed that 119 ESTs that were generated from this study had significant matches to unknown proteins and an additional 127 ESTs did not match any protein sequences in the NCBI database. Therefore, a motif search was carried out for the unknown ESTs to predict their putative functions. A total of 136 EST clusters were used in the motif analysis and the InterProScan software was chosen as the motif search tool. There were 66 types of motifs detected from this search. Based on the motifs detected within the query sequences, putative function predictions were successfully performed on 49 EST clusters.

Key words: *Calamus manan*, EST, open reading frames putative function prediction, motif detection, InterProScan

INTRODUCTION

The EST technology has been used for a few years now to analyse expressed genes within any given genome. This method has been used in studying differential gene expression, sex determination, developmental regulation, physiological functions, disease resistance and so much more in both monocot and dicot species (Torto-Alalibo *et al*., 2007). Functional analyses of ESTs that are generated from plants have been conducted in plants like rice, oil palm, barley and maize (Ho *et al*., 2007; Tan *et al*., 2005). Functional analysis that was conducted with the aid of computer softwares as well as experimental studies are useful in predicting or determining the function of ESTs generated from any library constructed (Ho *et al*., 2007; Lazo *et al*., 2004).

We constructed 8 cDNA libraries covering four developmental stages of flowering in the male and female flowers of *Calamus manan* Miq. *Calamus manan* Miq. or *Rotan manau* is a rattan species that produces large diameter canes. It has the highest commercial value in the furniture industry (Dransfield, 2001). As a dioecious palm, *C. manan* possesses both the male and female inflorescences on separate trees. So far, very little has been done in studying the molecular biology of *C. manan*. In our earlier research (Nadarajah *et al*., 2006, 2009), we had generated a set of ESTs from the male and female inflorescences cDNA libraries of *C. manan*. The purpose for this study was to identify flowering genes that could be used as probes to identify male and female *C. manan*. The flowering genes that are related to early flowering stage in development would be of preference as this will make the process of designing rattan plantations with the right male to female ratio possible. This is important as the right ratio would determine the success of generating sufficient rattan supply to sustain the industry.

The information obtained from the ESTs analysed provides a better understanding of the genes expressed in the *C. manan* inflorescence tissues. As previously described by Nadarajah *et al*. (2006) and Nadarajah *et al*. (2009), these ESTs were compared to the protein sequences deposited within NCBI database using BLASTX (Altschul *et al*., 1990) before characterisation into groups according to their protein functions. As a result of this characterisation, a group of ESTs that did not match any protein sequences in the NCBI database were identified. We classified these ESTs as *C. manan* novel sequences. In order to predict a putative function for this particular group of ESTs, a motif/domain analysis was carried out on the sequences to obtain sufficient information to suggest putative functions for these proteins (Lazo *et al*., 2004).

Corresponding Author: Dr. Kalaivani Nadarajah, School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

Motifs are small conserved regions found within protein sequences. They usually carry specific structural or functional significance. Detection of motifs among proteins with low sequence identities provide vital clues for the functional prediction of the proteins or to classify unknown proteins into functional families (Kawaoka *et al*., 2008). Although, the BLAST analysis provides identities for nucleotide and protein sequences, the classical BLASTN and BLASTX programmes within BLAST were unable to identify proteins with less than 30% identity (Rost and Valencia, 1996). Therefore, a motif detection system is used to assist in the prediction of function for the said protein sequence. Motif detection relies on the identification of small conserved regions in the proteins that can be identified in remote homologies, despite the lack of overall similarity (Hunger *et al*., 2003). Currently, motif structures are often used to predict putative functions for the query proteins based on the known functions of other proteins that share one or more motifs with the query protein. However, a query protein may exhibit more than one function if different motifs of the different functional families were found in that protein (Kawaoka *et al*., 2008; Tan *et al*., 2005).

In this study, two groups of *C. manan* floral ESTs from the earlier study (Nadarajah *et al*., 2006, 2009), containing ESTs that are either unknown or hypothetical proteins and ESTs that consists of novel protein sequences, were subjected to a motif search in order to predict the putative function of these ESTs based on matches to existing motifs deposited within the motif databases. The information obtained from the motif analysis will enable us to predict the putative functions of the unknown and hypothetical proteins in *C. manan* (Ho *et al*., 2007).

MATERIALS AND METHODS

The EST sequencing was conducted in the Genomic Laboratory in the Scottich Crop Research Institute (SCRI) in Scotland. The sequence annotation and analysis was conducted on the sequences from the ESTs from two libraries out of the eight libraries that were generated from *C. manan*.

ESTs selection and translation: Motif analysis was conducted on 93 assembled ESTs (15 contigs and 78 singletons) that were unknown and hypothetical proteins and 98 assembled ESTs (9 contigs and 89 singletons) that had no matches to any protein sequences in the NCBI database. In order to obtain useful information, only ESTs with sequence length of more than 180 bp were selected. Prior to motif search, these ESTs were translated to protein sequences using EMBOSS Transeq (http://www.ebi.ac.uk/Tools/emboss/transeq/index.html).

Selection of motif searching tool: InterPro database (http://www.ebi.ac.uk/interpro/ index.html) is a motif database that serves as the federation of various motif databases such as PROSITE (Hulo *et al*., 2006), Pfam (Finn *et al*., 2006), SMART (Letunic *et al*., 2006), TIGRFAMs (Haft *et al*., 2003), PRINTS (Attwood *et al*., 2003), ProDom (Bru *et al*., 2005), PIRSF (Wu *et al*., 2004), Gene3D (Yeats *et al*., 2006), Panther (Mi *et al*., 2005) and SuperFamily (Gough *et al*., 2001). The InterPro core is formed by merged annotations from the above mentioned member databases. Each combined InterPro entry includes functional descriptions and literature references. Extra information about patterns, profiles, fingerprints and so on can be obtained through the links made back to the relevant protein signatures from these member databases (Apweiler *et al*., 2001).

There are various types of motif data found in InterPro database such as protein families, domains, repeats, Post Translation Modification (PTM), binding sites and active sites. InterProScan (Zdobnov and Apweiler, 2001) is a tool that combines different protein recognition methods into one resource. Therefore, InterProScan was chosen as the motif search tool for use in this study. Each sequence was analysed through InterProScan to identify existing motifs within the architecture of the gene sequence. Motifs detected were used to predict a putative function for the gene.

RESULTS

Sequencing of cDNA and clustering of ESTs: As shown in the earlier study (Nadarajah *et al*., 2006), about 1529 good quality ESTs were obtained from the sequencing process of the male and female inflorescence libraries. Nine hundred and fifteen (915) of these ESTs (59.8%) were from the male inflorescence libraries and the remaining 614 ESTs (40.2%) were from the female inflorescence cDNA libraries. As a result of the clustering process, 229 contigs were assembled containing 805 ESTs (Table 1). The remaining 724 ESTs were singletons resulting in a redundancy of 24%.

Classification of ESTs according to protein function: The ESTs were then classified according to protein function through BLASTX analysis. The ESTs were classified into four categories. Figure 1 shows the breakdown of ESTs into the 4 categories. In this study, we report the

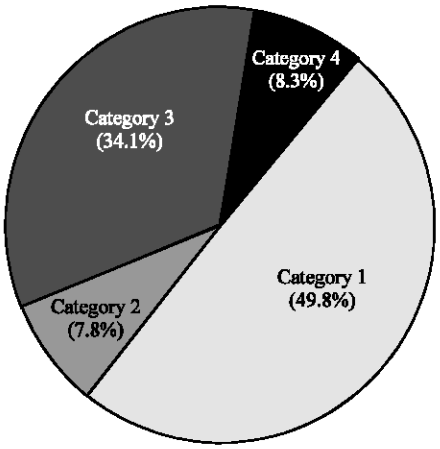

Fig. 1: 1529 ESTs from *C. manan* inflorescence libraries categorised into 4 groups: Category 1: Significant with known function, Category 2: Significant with unknown function and hypothetical proteins, Category 3: Not significant and Category 4: No match

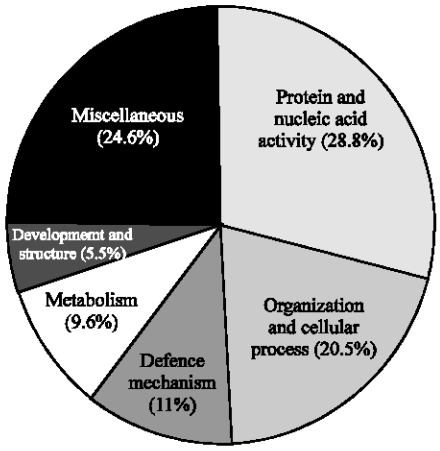

Fig. 2: Categorisation of a set of *C. manan* floral ESTs into 6 functional classes based on motifs' functions

Table 1: Summary of sequencing and clustering of *C. manan* floral ESTs

Output	Amount
cDNA clones sequenced	2688
ESTs generated	2063
Good quality ESTs	1529
Average length of EST (bp)	400
Contig assembly	
No. of contigs	229
No. of singletons	724
No. of sequence clusters	953

characterisation of category 2 and 4 ESTs as ESTs in these categories have not been functionally classified. One hundred and ninety one (191) assembled ESTs (24 contigs and 167 singletons) from these two categories were subjected to a motif scan to help predict a putative function in order to identify any novel proteins or flowering related ESTs that could be studied further in the future.

Selection of ESTs and motif search: The ESTs with length less than 180 bp were discarded, leaving behind a total of 143 assembled sequences (with an average length of 180-752 bp) for translation. The translated ORFs that ranged from 20 amino acids and above were selected manually for motif search via the InterProScan software.

Motifs and putative functions: The results of InterProScan motif search showed that 7 assembled ESTs had no motifs. The remaining 135 assembled ESTs matched 105 types of protein motifs in the InterProScan database. There were two types of motifs found abundantly in these ESTs. They were the signal peptides and trans-membrane helices. Almost 91.9% (124 clusters) of these ESTs contained at least a single signal peptide motif. This motif was detected through SignalP 3.0 server. SignalP 3.0 server predicts the presence and location of the signal peptide cleavage sites in amino acid sequences from different organisms based on a combination of several artificial neural networks and the Hidden Markov Model (Bendtsen *et al*., 2004). Seventy four clusters (54.8%) contained the trans-membrane helices motif. This motif was detected though TMHMM Server v.2.0, which predicts the presence of the trans-membrane helices based on the Hidden Markov model (Krogh *et al*., 2001). Since, these two motifs are common fixtures in the genes architecture, they were therefore excluded from this analysis. In total there were 70 EST clusters that contained only these two motifs. The remaining 65 assembled ESTs and their putative functions were predicted based on the motif(s) found on their sequences (Table 2). There were 34 assembled sequences that matched a single motif (excluding the signal peptide and trans-membrane helices motifs) and 31 clusters that had more than one motif.

Functional classification of motifs and ESTs: Based on the respective putative motif function, these assembled ESTs were briefly categorised into several main functional classes, which were protein and nucleic acid activity, metabolism, organisation and cellular process, defence mechanism, development and structure and miscellaneous. This classification was based on the functional categories assigned to *C. manan* floral ESTs in the previous study (Nadarajah *et al*., 2006) with some modification (Table 2). The percentage of assembled ESTs with assigned functional classes is shown in Fig. 2.

Table 2: The putative functions for the motifs detected within the assembled ESTs of *C. manan*

AssembledEST/ functional class	Motif name	Motif sequence	Motif location	Putative function
Protein and nucleic acid activity				
Contig 3[o,c]	IPR007808 Protein of unknown function DUF701, zinc-binding putative	-	3-89aa of ORF1a (89aa)	Unknown (zinc binding)
	PF05129 Transcription elongation factor Elf1 like	-	9-69aa of ORF1a (89aa)	Transcription elongation factor
Contig 18	IPR002775 Alba, DNA/RNA-binding protein	-	1-61aa of ORF2 (102aa)	Maintain structural and functional stability of RNA and ribosomes
	IPR014560 Uncharacterised conserved protein UCP030333, DNA/RNA-binding Alba-related	-	2-102aa of ORF2 (102aa)	RNA and DNA binding
Contig 99	IPR003107 RNA-processing protein, HAT helix	-	45-77, 80-112 and 115-147aa of ORF3 (181aa)	RNA processing
	IPR011990 Tetratricopeptide-like helical	-	45-153aa of ORF3 (181aa)	Binding activity
	IPR013026 Tetratricopeptide region	-	45-134aa of ORF3 (181aa)	Mediates protein-protein interactions
ECM01.GS001.D05	IPR007087 Zinc finger, C2H2-type	C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H	56-78aa of ORF1 (206aa)	Zinc ion and nucleic acid binding in gene transcription
ECM01.GS001.J03[c]	IPR005819 Histone H5	-	33-47, 68-85 and 90-109aa of ORF6 (111aa)	Gene regulation
ECM01.GS010.C07	IPR002733 AMMECR1	LRGCIG (conserved domain)	1-65aa of ORF1 (82aa)	Transcription, replication, repair or translation machinery
ECM01.GS010.E07	IPR000571 Zinc finger, CCCH-type	C-x(8)-C-x(5)-C-x(3)-H	5-81 and 122-148aa of ORF2 (170aa)	Regulation of transcription; involved in cell cycle
	PTHR10288 RNA-binding protein related	-	41-97aa of ORF2 (170aa)	RNA binding
ECM01.GS010.L12[o]	IPR007757 MT-A70	-	2-26aa of ORF2 (177aa)	Methyltransferase activity in nucleotide and nucleic acid metabolic process
	PTHR13107 Karyogamy protein KAR4-related	-	1-121aa of ORF2 (177aa)	Transcription factor
	IPR000408 Regulator of chromosome condensation, RCC1	[LIVMFA]-[STAGC](2)-G-x-{TAV}-H-[STAGLI]-[LIVMFA]-{KI}-[LIVM]	28-38aa of ORF3 (46aa)	Guanine-nucleotide dissociation stimulator; regulate gene expression
ECM01.GS010.M12	IPR001487 Bromodomain	[STANVFHG]-x(2)-[FAS]-x(4)-[DNSPAKT]-x(0,7)-[DENQTFG]-Y-[HFYLRKT]-x(2)-[LIVMFY]-x(3)-[LIVM]-x(4)-[LIVM]-x(6,10)-Y-x(12,13)-[LIVM]-x(2)-N-[SACF]-x(2)-[FY]	42-55, 56-72, 72-90 and 90-109aa of ORF2 (143aa)	Transcriptional activation
	PTHR13900 Transcription initiation factor TFIID	-	3-135aa of ORF2 (143aa)	Transcription process
ECF01.GS011.C17	IPR001841 Zinc finger, RING-type	C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVMYA]	84-126aa of ORF3 (207aa)	Zinc ion and nucleic acid binding; E3 ubiquitin-protein ligase activity
	IPR013083 Zinc finger, RING/FYVE/PHD-type	C-x(1,2)-C-x(5,45)-[VMFLWIE]-x-C-x(1,4)-C-x(1,4)-[WYFVQHLT]-C-xC-x(5,45)-[WFLYI]-x-C-x(2)-C	77-148aa of ORF3 (207aa)	Zinc ion and nucleic acid binding
	PTHR12456 Topoisomerase 1-binding RING finger-related	C-X(2)-C-X(9,39)-C-X(1,3)-H-X(2,3)-[NCH]-X(2)-C-X(4,48)-C-X(2)-C.	44-140aa of ORF3 (207aa)	protein-protein interactions in regulation of transcription
	SSF57850 RING/U-box	-	77-148aa of ORF3 (207aa)	DNA replication/repair
ECF01.GS011.N17	IPR000504 RNA recognition motif, RNP-1	-	27-98aa of ORF1 (115aa)	nucleic acid binding
	SSF54928 RNA-binding domain, RBD	-	45-115aa of ORF1 (115aa)	RNA binding, metabolism and transport
ECF02.GS010.A19[o,c]	IPR001163 Like-Sm ribonucleoprotein, core	-	12-89aa of ORF1 (110aa)	modulators of RNA biogenesis and function; translation
	IPR006649 Like-Sm ribonucleoprotein, eukaryotic and archaea-type, core	-	22-89aa of ORF1 (110aa)	modulators of RNA biogenesis and function; translation
	IPR010920 Like-Sm ribonucleoprotein-related, core	-	20-110aa of ORF1 (110aa)	Modulators of RNA biogenesis and function; translation
	IPR016654 U6 snRNA-associated Sm-like protein Lsm2	-	18-110aa of ORF1 (110aa)	Modulators of RNA biogenesis and function; translation

Table 2: Continued

AssembledEST/ functional class	Motif name	Motif sequence	Motif location	Putative function
ECF02.GS010.A22	IPR000504 RNA recognition motif, RNP-1	-	80-150aa of ORF2 (194aa)	Nucleic acid binding
	IPR012677 Nucleotide-binding, alpha-beta plait	-	78-153aa of ORF2 (194aa)	Nucleotide binding
	SSF54928 RNA-binding domain, RBD	-	41-193aa of ORF2 (194aa)	RNA binding, metabolism and transport
ECF02.GS010.D19	IPR005326 Plectin/S10, N-terminal	-	5-100aa of ORF3 (173aa)	RNA binding (found in ribosomal S10 protein)
	PTHR12146 40S Ribosomal protein S10	-	4-169aa of ORF3 (173aa)	Translation
ECF02.GS010.H24	IPR006456 ZF-HD homeobox protein Cys/His-rich dimerisation region	-	68-127aa of ORF1 (131aa)	Transcription factor activity
ECF02.GS010.O19	PTHR10052:SF2 60S Ribosomal protein L18A, plant	-	1-148aa of ORF2 (158aa)	Translation
EcF03.GS002B.H02	IPR000637 HMG-I and HMG-Y, DNA-binding	[AT]-x(1,2)-[RK](2)-[GP]-R-G-R-P-[RK]-x	48-60aa of ORF6 (83aa)	nucleosome phasing; regulation of transcription
ECM04.GS001 A.C01[o, c]	IPR000637 HMG-I and HMG-Y, DNA-binding	-	51-61, 96-107 and 110-120aa of ORF1 (131aa)	nucleosome phasing; regulation of transcription
	IPR000976 Wilm's tumour protein	-	60-76 and 94-108aa of ORF3 (130aa)	transcription factor activity
ECM04.GS001A.F02	IPR005345 PHF5-like	Contain five CXXC motifs	3-78a of ORF4 (82aa)	transcription regulation
ECM04.GS002C.H06	SSF54060 His-Me finger endonucleases	-	63-158aa of ORF4 (185aa)	DNA replication/repair
ECM04.GS002D.C11	IPR001841 Zinc finger, RING-type	C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVMYA]	15-24aa of ORF1 (24aa)	Zinc ion and nucleic acid binding; E3 ubiquitin-protein ligase activity
Subtotal = 21 (28.8%)				
Organisation and cellular process				
Contig 106	IPR002890 Alpha-2-macroglobulin, N-terminal	-	2-185aa of ORF6 (198aa)	Endopeptidase inhibitor activity
Contig 157	IPR007717 NPL4	-	5-61aa of ORF2 (64aa)	Nuclear transport
Contig 225	IPR000644 Cystathionine beta-synthase (CBS), core	-	115-166aa of ORF1 (173aa)	Intracellular targeting and trafficking; sensor of intracellular metabolites
ECM01.GS001.E06[o]	IPR000886 Endoplasmic reticulum, targeting sequence	[KRHQSA]-[DENQ]-E-L>	142-145aa of ORF3 (145aa)	Prevent secretion from ER, signal transduction at endoplasmic reticulum
	IPR005829 Sugar transporter, conserved site	-	97-113aa of ORF6 (132aa)	Transport of various carbohydrates, organic alcohols and acids
ECM01.GS001.J03[c]	IPR003993 Treacher Collins syndrome, treacle	-	2-20 and 41-64aa of ORF6 (111aa)	Nucleolar trafficking protein
ECM01.GS010.N11	IPR001638 Bacterial extracellular solute-binding protein, family 3	-	104-117aa of ORF3 (159aa)	Active transport of solutes across the cytoplasmic membrane
ECF01.GS011.F17[o]	IPR001024 Lipoxygenase, LH2	-	65-192aa of ORF3 (200aa)	Mediate membrane attachment via other protein binding partners
	IPR008976 Lipase/ lipooxygenase, PLAT/LH2	-	65-181aa of ORF3 (200aa)	Mediate membrane attachment via other protein binding partners
	IPR010916 TonB box, conserved site	-	1-18aa of ORF5 (47aa)	Involved in the interaction of the protein with the tonB protein
ECF02.GS010.A19[o, c]	IPR000194 ATPase, F1/V1/A1 complex, alpha/beta subunit, nucleotide-binding	P-[SAP]-[LIV]-[DNH]-{LKGN}-{F}-{S}-S-{DCPH}-S	12-21aa of ORF2 (37aa)	Catalyzing transmembrane movement of substances
ECF02.GS010.G23	IPR002048 Calcium-binding EF-hand	D-{W}-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x(2)-[DE]-[LIVMFYW]	36-67 and 75-101aa of ORF3 (101aa)	Calsium ion binding
	IPR011992 EF-Hand type	-	30-100aa of ORF3 (101aa)	Calsium ion binding
	PTHR10891 CALMODULIN	-	32-101aa of ORF3 (101aa)	Calsium ion binding
ECF02.GS010.H22	IPR011042 Six-bladed beta-propeller, TolB-like	-	1-153aa of ORF3 (188aa)	Mediate protein-protein interactions with colicins
	IPR011659 WD40-like Beta Propeller / PD40	-	2-59aa of ORF3 (188aa)	Involved in various functions, e.g., signal transduction, transcription regulation, cell cycle control and apoptosis
	SSF50960 TolB, C-terminal domain	-	1-156aa of ORF3 (188aa)	Mediate protein-protein interactions with colicins
ECF02.GS010.K23[o]	IPR000073 Alpha/beta hydrolase fold-1	-	93-174aa of ORF2 (209aa)	Hydrolase activity

Table 2: Continued

AssembledEST/ functional class	Motif name	Motif sequence	Motif location	Putative function
	PTHR15913 acid cluster protein 33	-	20-208aa of ORF2 (209aa)	Intracellular transportation, receptor-mediated signalling pathway
	IPR000215 Protease inhibitor I4, serpin		21-31aa of ORF3 (62aa)	Irreversible serine protease inhibitors
ECF02.GS010.O21	IPR007233 Sybindin-like protein	-	43-161aa of ORF1 (161aa)	Involved in endoplasmic reticulum-to-Golgi vesicle transport
	IPR011012 Longin-like /SNARE-like	-	41-161aa of ORF1 (161aa)	Involved in endoplasmic reticulum-to-Golgi vesicle transport
	PTHR23249:SF9 Trafficking protein particle complex subunit 1	-	50-161aa of ORF1 (161aa)	Intracellular protein traffic
ECM04.GS001C.D05	IPR002345 Lipocalin	[DENG]-{A}-[DENQGSTARK]-x(0,2)-[DENQARK]-[LIVFY]-{CP}-G-{C}-W-[FYWLRH]-{D}-[LIVMTA]	23-34aa of ORF3 (39aa)	Transport of nutrients and pheromone, control of cell regulation
ECM04.GS001C.D06	IPR000804 Clathrin adaptor,sigma subunit/coatomer, zeta subunit	[LIVMC]-[LIVM]-Y-[KR]-x(4)-L-Y-F	12-22aa of ORF2 (31aa)	Protein sorting in the late-Golgi/trans-Golgi network
ECM04.GS001D.C08	IPR000644 Cystathionine beta-synthase (CBS), core	-	118-167aa of ORF3 (178aa)	Intracellular targeting and trafficking; sensor of intracellular metabolites
Subtotal = 15 (20.5%)				
Defence mechanism				
ECM01.GS001.F02	IPR002579 Methionine sulphoxide reductase B	-	19-144aa of ORF3 (144aa)	Reductase activity; protect against oxidative damage
	IPR011057 Mss4-like	-	17-144aa of ORF3 (144aa)	Signal transduction
ECM01.GS010.E08[b,c]	IPR006121 Heavy metal transport/detoxification protein	[LIVNS]-x-{L}-[LIVMFA]-x-C-x-[STAGCDNH]-C-x(3)-[LIVFG]-{LV}-x(2)-[LIV]-x(9,11)-[IVA]-x-[LVFYS]	30-90aa of ORF2 (157aa)	Involved in bacterial resistance to toxic metals, e.g., cadmium and lead
	PTHR22814 Copper transport protein ATOX1-related	-	34-157aa of ORF2 (157aa)	Copper transport; involved in cellular antioxidant defence
ECM01.GS010.J07	PTHR10992:SF13 Sigma factor SIGB regulation protein RSBQ	-	1-68aa of ORF1 (73aa)	Immunity and defence; regulate energy stress activation of the sigma-B transcription factor
ECM01.GS010.J07	SSF53474 alpha/beta-Hydrolases	-	1-68aa of ORF1 (73aa)	Hydrolase activity
ECF01.GS011.E14[c] stress protein (Usp) A	IPR006015 Universal	-	22-40, 132-144 and 150-172aa of ORF3 (173aa)	Stress endurance activity
	IPR006016 UspA	-	22-172aa of ORF3 (173aa)	Stress endurance activity
	IPR014729 Rossmann-like alpha/beta/alpha sandwich fold	-	25-172a of ORF3 (173aa)	Unknown (found in a lot of proteins including UspA)
EcF03.GS001A.A04	IPR000871 / PS00146 Beta-lactamase, class A	[FY]-x-[LIVMFY]-{E}-S-[TV]-x-K-x(3)-{T}-[AGLM]-{D}-{KA}-[LC]	14-29aa of ORF3 (30aa)	Beta-lactam antibiotic catabolic process
ECF02.GS010.E23	IPR006121 Heavy metal transport/detoxification protein	[LIVNS]-x-{L}-[LIVMFA]-x-C-x-[STAGCDNH]-C-x(3)-[LIVFG]-{LV}-x(2)-[LIV]-x(9,11)-[IVA]-x-[LVFYS]	29-96aa of ORF2 (157aa)	Involved in bacterial resistance to toxic metals, e.g., cadmium and lead
	PTHR22814 Copper transport protein ATOX1-related	-	35-157aa of ORF2 (157aa)	Copper transport; involved in cellular antioxidant defence
ECF03.GS002D.B08	IPR006121 Heavy metal transport/detoxification protein	[LIVNS]-x-{L}-[LIVMFA]-x-C-x-[STAGCDNH]-C-x(3)-[LIVFG]-{LV}-x(2)-[LIV]-x(9,11)-[IVA]-x-[LVFYS]	10-74aa of ORF1 (83aa)	Involved in bacterial resistance to toxic metals, e.g., cadmium and lead
	PTHR22814 Copper transport protein ATOX1-related	-	23-83aa of ORF1 (83aa)	Copper transport; involved in cellular antioxidant defence
ECM04. GS002D.G12[b,c]	IPR007138 Antibiotic biosynthesis monooxygenase	-	36-99aa of ORF1 (120aa)	Antibiotic biosynthesis
	IPR011008 Dimeric alpha-beta barrel	-	23-120aa of ORF1 (120aa)	Unknown
Subtotal = 8 (11.0%)				
Metabolism				
ECM01.GS001.G01	PTHR13902 Serine/threonine-protein kinase WNK (With No Lysine)-related	-	3-65aa of ORF3 (193aa)	Protein kinase activity, ATP binding
ECM01.GS001.L01[b,c]	IPR017441 Protein kinase ATP binding, conserved site	[LIV]-G-{P}-G-{P}-[FYWMGSTNH]-[SGA]-{PW}-[LIVCAT]-{PD}-x-[GSTACLIVMFY]-x(5,18)-[LIVMFYWCSTAR]-[AIVP]-[LIVMFAGCKR]-K	48-74aa of ORF4 (76aa)	Protein kinase activity, ATP binding

Table 2: Continued

AssembledEST/ functional class	Motif name	Motif sequence	Motif location	Putative function
ECM01.GS010.G07	IPR000868 Isochorismatase hydrolase	-	1-105aa of ORF2 (111aa)	Catalyses the conversion of isochorismate
	PTHR11080 Pyrazinamidase/ Nicotinamidase	-	8-109aa of ORF2 (111aa)	Nicotinamidase activity
ECF01.GS011.A14	IPR007941 Protein of unknown function DUF726	-	13-187aa of ORF2 (203aa)	Unknown
	SSF53474 alpha/beta-Hydrolases	-	56-157aa of ORF2 (203aa)	Hydrolase activity
ECF01.GS011.E14[c]	SSF52402 Adenine nucleotide alpha hydrolases-like	-	25-172aa of ORF3 (173aa)	Nucleotide metabolism
ECM04.GS002C.A07	IPR000160 GGDEF	GG[DE][DE]F (conserved central sequence pattern)	80-231aa of ORF1 (242aa)	Diguanylate cyclase activity; regulate exopolysaccharide synthesis
ECM04. GS002D.G12[o,c]	IPR000583 Glutamine amidotransferase, class-II	-	1-14aa of ORF5 (35aa)	Biosynthesis of purine and asparagine
Subtotal = 7 (9.6%)				
Development and structure				
Contig 3[o]	IPR001545 Gonadotropin, beta chain	C-[STAGM]-G-[HFYL]-C-x-[ST]	5-11aa of ORF1b (22aa)	Hormone activity
Contig 224	IPR013919 Peroxisome membrane protein, Pex16	-	1-135aa of ORF2 (136aa)	Membrane protein
ECM01.GS001.L01[o,c]	IPR008801 Rapid ALkalinisation Factor (RALF)	-	1-40aa of ORF2 (40aa)	Arrests root growth and development
ECM01.GS010.D10	IPR004345 TB2/DP1 and HVA22 related protein	-	23-76aa of ORF3 (76aa)	Regulate abscisic acid-inducing
Subtotal = 4 (5.5%)				
Miscellaneous				
Contig 159	IPR007493 Protein of unknown function DUF538	-	3-80aa of ORF1 (102aa)	Unknown
Contig 215	IPR007608 Protein of unknown function DUF584	-	2-112aa of ORF3 (113aa)	Unknown
ECM01.GS001.L06	IPR008579 Protein of unknown function DUF861, cupin-3	-	80-155aa of ORF2 (158aa)	Unknown
	IPR011051 Cupin, RmlC-type	-	68-157aa of ORF2 (158aa)	Unknown
	IPR014710 RmlC-like jelly roll fold	-	56-157aa of ORF2 (158aa)	Unknown
ECM01.GS001.P02	IPR007462 Protein of unknown function DUF502	-	112-213aa of ORF2 (213aa)	Unknown
ECM01.GS010.A12	IPR005061 Protein of unknown function DUF292, eukaryotic	-	50-126aa of ORF3 (126aa)	Unknown
ECM01.GS010.C08	IPR006702 Protein of unknown function DUF588	-	6-42aa of ORF3 (77aa)	Unknown
ECM01.GS010.E08[o,c]	IPR013032 EGF-like region, conserved site	C-x-C-x(2)-{V}-x(2)-G-{C}-x-C	17-28aa of ORF6 (46aa)	Unknown
ECF01.GS011.L18	IPR013032 EGF-like region, conserved site	C-x-C-x(2)-{V}-x(2)-G-{C}-x-C	2-13aa of ORF5 (63aa)	Unknown (found in the sequence of epidermal growth factor)
ECF02.GS010.L20	IPR009606 Protein of unknown function DUF1218	-	32-147aa of ORF2 (173aa)	Unknown
ECF03.GS001C.A06	IPR002876 Protein of unknown function DUF28	-	1-37aa of ORF3 (45aa)	Unknown
	SSF75625 YebC-like	-	1-40aa of ORF3 (45aa)	Unknown
ECM04.GS001A.C01[o,c]	PTHR10499 Collagen alpha chain	-	88-129aa of ORF5 (131aa)	unknown
ECM04.GS001B.D12	IPR006461 Protein of unknown function Cys-rich	-	10-30aa of ORF6 (59aa)	Unknown
ECM04.GS001D.A05	IPR007432 Protein of unknown function DUF480	-	23-126aa of ORF1 (126aa)	Unknown
ECM04.GS001D.C09	PTHR13495 NEFA-interacting nuclear protein NIP30	-	4-84aa of ORF3 (145aa)	Unknown
ECM04.GS001D.C10	PTHR12447 Uncharacterised with Ankyrin repeat domain	-	3-103aa of ORF3 (108aa)	Protein-protein interaction domains
ECM04.GS002C.H08	IPR009606 Protein of unknown function DUF1218	-	55-170aa of ORF3 (197aa)	Unknown
ECM04.GS002D.C02	IPR007897 PHB accumulation regulatory	-	26-66 and 76-116aa of ORF3 (133aa)	Regulating polymer accumulation
ECM04.GS005D.A01	IPR006946 Protein of unknown function DUF642	-	2-177a of ORF3 (177aa)	Unknown
Subtotal = 18 (24.6%)				
Total = 73				

[o]Assembled EST with more than one ORF, [c]Assembled EST which had been categorised into two functional classes. aa: Amino acids, motif sequences are not available on InterProScan for all motifs

For ESTs with motifs that were related to ribosomal proteins, histones and all sorts of transcription or translation activities (including nucleic acid metabolism, binding and protein ubiquitination) were included in the functional class of protein and nucleic acid activity. As shown in Table 2, there were 21 assembled ESTs grouped into this class. Four of them, namely Contig 3, ECM01.GS001.J03, ECF02.GS010.A19 and ECM04.GS001A.C01, were also members of other functional classes as they contained different types of motifs (different functions) either within the same translated frame or in different ORFs. Nine EST clusters showed the presence of single motifs, while 12 EST clusters had their functions predicted based on combination of several motifs present within their ORFs (Table 2).

Motifs showing putative functions of membrane protein interaction, binding and transportation of cellular molecules, cellular activity regulation and inhibition and signal transduction were grouped into the organisation and cellular process functional class, that consists of 15 assembled ESTs. Among these sequences, 9 showed single motifs while 6 others showed the presence of two to three motifs. Contig 225 and ECM04.GS001D.C08 had the Cystathionine Beta-Synthase (CBS), core motif.

The EST clusters that contained more than one motif were assigned their functions based on the presence of their motifs at their single or multiple translated frames. Although, ECF02.GS010.G23 and ECF02.GS010.O21 had three motifs each found on their single translated frames, their functions were easily given because their motifs had the same or very similar functions, as shown in Table 2.

There were two motifs found in ECM01.GS001.E06, i.e., endoplasmic reticulum, targeting sequence at ORF3 and sugar transporter, conserved site at ORF6. The lipoxygenase, LH2 and lipase/lipooxygenase, PLAT/LH2 motifs were detected in ORF3 of ECF01.GS011.F17. The ORF2 of ECF02.GS010.K23 contained motifs alpha/beta hydrolase fold-1 and acid cluster protein 33, while motif proteinase inhibitor I4 serpin was found on the ORF3 of this EST.

As shown in Table 2, the functional class for defence mechanism consists of motifs with functions related to defence and resistance, regulation of environmental stress, or detoxification process. ECF03.GS001A.A04 was the only EST with a single motif: β-lactamase, class A, while the other 6 ESTs had two or three motifs detected within the same or different ORFs. β-lactamase is involved in the hydrolysis of the β-lactam ring of β-lactam antibiotics, such as penicillins and cephalosporins.

Four ESTs (ECM01.GS001.F02, ECF01.GS011.E14, ECM01.GS010.J07 and ECM04.GS002D.G12) functions were predicted based on the combination of motifs found within the same translated frame of each EST. Motifs methionine sulphoxide reductase B and Mss4-like were detected on ORF3 of ECM01.GS001.F02 (Table 2). ECM01.GS010.J07 had been assigned similar function to the RSBQ motif. The ORF1 of ECM04.GS002D.G12 contained motifs for antibiotic biosynthesis monooxygenase and dimeric alpha-beta barrel. The dimeric alpha-beta barrel motif's function remains unknown and it is found in various proteins, including bacterial actinorhodin biosynthesis monooxygenase (Sciara *et al*., 2003). Thus, this EST was likely to be involved in antibiotic biosynthesis. As mentioned earlier, ECM04.GS002D.G12 was also grouped into the metabolism functional class of motifs (Table 2).

All the motifs that exhibited functions related to catalytic or energy (ATP) related activity were categorised into the metabolism functional class. There were three ESTs that were categorised into two different functional classes, i.e., ECM01.GS001.L01, ECF01.GS011.E14 and ECM04.GS002D.G12. ECF01.GS011.E14 has different motifs that were detected within the same frame (ORF3). They were the adenine nucleotide alpha hydrolases-like and other motifs (mentioned later) that belong to the defence mechanism class of proteins. On the other hand, both ECM01.GS001.L01 and ECM04.GS002D.G12 showed two types of motifs within two different translated frames. ORF4 of ECM01.GS001.L01 contained the motif protein kinase ATP binding conserved site, while the ORF5 of ECM04.GS002D.G12 had the glutamine amidotransferase, class-II. The other translated frames of these two ESTs had motifs that belonged to the development and structure and defence functional classes, respectively. For ECF01.GS011.A14 and ECM01.GS010.G07, two motifs were detected within their translated frames, respectively. ECF01.GS011.A14 has the alpha/beta-hydrolases motif with an unknown function. The two motifs within ECM01.GS010.G07 have catalytic function. The remaining two ESTs in this functional class showed the presence of a single motif within their translated frames (Table 2).

The development and structure functional class was made up of assembled ESTs that contained motifs of structural importance or were involved in hormone activity. There were four EST clusters and each of them had only one motif. Among them, two clusters had been categorised into other functional classes as well, i.e., Contig 3 and ECM01.GS001.L01. Contig 3 contained the gonadotropin, β-chain motif in its translated ORF1b frame (Table 2).

Lastly, there were 18 assembled ESTs categorised into the miscellaneous functional class. The motifs that belonged to this class usually didn't exhibit any distinct

function or had unknown function. Examples of these motifs are cupin, RmlC-type, RmlC-like jelly roll fold, EGF-like region, conserved site, NEFA-interacting nuclear protein NIP30, YebC-like, collagen α-chain, PHB accumulation regulatory and uncharacterised Ankyrin repeat domain. As mentioned earlier, ECM01.GS010.E08 and ECM04.GS001A.C01 were also members of defence mechanism and protein and nucleic acid activity functional classes, respectively. EGF-like region, conserved site was detected within ORF6 of ECM01.GS010.E08. This EGF domain was found in the sequence of epidermal growth factor and the actual function of this motif is unknown (Downing *et al*., 1996). The collagen α-chain which had no known function, was found at ORF5 of ECM04.GS001A.C01 (Table 2).

The overall results showed that 49 EST clusters obtained from this analysis have 81 motifs with known functions. However, there were no floral-specific motifs detected in this study. According to Fig. 2, most of the ESTs generated in this study were found to play an important role in various protein and nucleic acid activities. This was followed by ESTs that showed functions in organisation and cellular processes, defence mechanism and metabolism. Through the ESTs analysed in this study the least number of ESTs were found involved in developmental and structural processes. In addition, almost one quarter of the ESTs had shown similarity to motifs of unknown function.

DISCUSSION

Motif search was conducted to predict the putative function of ESTs through motif detection from a set of *C. manan* floral ESTs that either showed significant matches to unknown proteins, or did not have matches to any proteins in NCBI database. In this study, InterProScan was selected as a motif searching tool. The assembled ESTs were assigned their putative functions based on the motifs found in their sequence architecture and then categorised into 6 main functional classes. Therefore, a putative function for each EST was made based on the motif(s) present in the architecture of their sequences. However, no functions were predicted for sequences with motifs of unknown function. These putative functions are merely preliminary predictions and further analysis is needed to certify their functions. The results of the preliminary predictions are discussed below.

The ESTs that contain a single motif were assigned a functional class easily by looking at the function exhibited by the motif. However for ESTs with more than one motif, they were grouped into functional classes according to the collective prediction on the function of motifs that are present. For example, ECF02.GS010.A22 matched RNA recognition motif, RNP-1, RNA-binding domain, RBD and nucleotide-binding, alpha-beta plait motifs. Many eukaryotic proteins contain a putative RNA-binding domain which is known to bind single-stranded RNAs (Dreyfuss *et al*., 1988). The largest group of single stranded RNA-binding proteins was the eukaryotic RNA Recognition Motif (RRM) family that contains a RNP-1 consensus sequence (Bandziulis *et al*., 1989; Query *et al*., 1989). RRM proteins have a variety of RNA binding preferences and function as a regulator of alternative splicing, RNA stability and translation (Sachs *et al*., 1987; Query *et al*., 1989). Nucleotide-binding, alpha-beta plait is a domain found in various RNA-binding or DNA-binding domains (Kielkopf *et al*., 2004) or in the ribosomal protein L23 (Chenuil *et al*., 1997). Therefore, it can be predicted that this EST may be involved in transcription and translation activities.

Contig 3 had similarity to motifs protein of unknown function DUF701, zinc-binding putative and transcription elongation factor Elf1 like. The first motif is a zinc binding domain without significant function, while the second motif has been identified as a transcription elongation factor in *Saccharomyces cerevisiae* (Prather *et al*., 2005). Thus, this EST cluster was predicted to be involved in transcription activity based on the information from the second motif and that zinc fingers are structures found in the regulatory protein such as those involved in the process of transcription and translation.

Contig 99 contains RNA-processing protein, HAT helix, tetratricopeptide-like helical, tetratricopeptide region within the ORF3 of the translated EST. According to Preker and Keller (1998), HAT containing proteins were components of macromolecular complexes which were required for RNA processing. Tetratricopeptide-like helical is a multi-helical fold domain which can bind ligands at many different regions and it has multiple functional roles (Andrade *et al*., 2001). The Tetratricopeptide Repeat Region (TPR) is a motif found in a wide range of proteins (Goebl and Yanagida, 1991), that mediate protein-protein interactions and the assembly of multiprotein complexes (D'andrea and Regan, 2003). Proteins containing TPRs were involved in a variety of biological processes, such as cell cycle regulation, transcriptional control, mitochondrial and peroxisomal protein transport. Both RNA-processing protein, HAT helix and tetratricopeptide region motifs were found within tetratricopeptide-like helical domain. Although there were various functions contributed by the tetratricopeptide motifs, Contig 99 was predicted to have similar function as a RNA-processing protein, HAT helix motif, therefore placing this contig in the RNA processing functional class.

There were several types of zinc finger domains found in this analysis: C2H2-type, RING-type, RING/FYVE/PHD-type, RING/U-box and CCCH-type. Zinc finger (Znf) domains were relatively small protein motifs that can bind zinc atoms. They were first identified as a DNA-binding motif in transcription factor TFIIIA from *Xenopus laevis*, but now they were recognised to bind DNA, RNA, protein and lipid substrates (Klug, 1999; Matthews and Sunde, 2002). It was suggested that Znf motifs had evolved specialised functions, such as gene transcription, translation, mRNA trafficking, cytoskeleton organisation, epithelial development, cell adhesion, protein folding, chromatin remodelling and zinc sensing (Laity *et al*., 2001). C2H2-type Znfs were the first class to be characterised and they are the most common DNA-binding motifs found in eukaryotic transcription factors (Bouhouche *et al*., 2000). The RING-type is a protein interaction domain involved in a diverse range of biological processes. E3 ubiquitin-protein ligase activity was likely to be the general function of the RING domain (Freemont, 1993; Borden and Freemont, 1996). RING/FYVE/PHD-type and RING/U-box were similar to RING zinc finger. According to Wang *et al*. (2008), the zinc finger CCCH-type was involved in regulation of transcription and, it also played an important role in cell cycle or growth phase-related regulation.

Contig 225 and ECM04.GS001D.C08 contained CBS domains which act to bind adenosine derivatives (Scott *et al*., 2004). This domain is involved in the intracellular targeting and trafficking in the chloride ion channels (Carr *et al*., 2003) and is also involved as a sensor of intracellular metabolites (Scott *et al*., 2004). The motif found at ORF6 of ECM01.GS001.J03 was Treacher Collins syndrome, treacle. It is common to nucleolar trafficking proteins which contain putative nuclear and nucleolar localisation signals (Wise *et al*., 1997). The motif ATPase, F1/V1/A1 complex, alpha/beta subunit, nucleotide-binding was found in ORF2 of ECF02.GS010.A19. This motif is involved in ATP synthesis coupled with proton transport and catalyzing trans-membrane movement of substances. Both these ESTs were also categorised into protein and nucleic acid activity functional class, due to the different motifs present at either the same frame (ECM01.GS001.J03) or different frame (ORF1 of ECF02.GS010.A19) that reflected this function.

ECF02.GS010.G23 is predicted to be involved in calcium ion binding, while ECF02.GS010.O21 played a role in cellular transport. As for ECF02.GS010.H22, the ORF3 showed matches to motifs 6-bladed beta-propeller, TolB-like, TolB, C-terminal domain and WD40-like Beta Propeller/PD40. The first two motifs are involved in mediating protein-protein interactions with colicins (Carr *et al*., 2000), while the latter is involved in various functions such as signal transduction, transcription regulation, cell cycle control and apoptosis (Li and Roberts, 2001). The WD40-like beta propeller can be found within 6-bladed beta-propeller, TolB-like motif. Thus, this EST may be involved in regulation of protein-protein interactions in plant defence.

The sugar transporters found in ECM01.GS001.E06 belonged to a membrane proteins superfamily which is involved in the binding and transport of various carbohydrates, organic alcohols and acids (Mueckler *et al*., 1985). On the other hand, the endoplasmic reticulum, targeting sequence was expected to play a role in signal transduction at the endoplasmic reticulum (ER) and prevent secretion from ER (Munro and Pelham, 1987; Pelham, 1990). So, this EST can be a membrane protein involved in the signalling and transportation processes.

The lipoxygenase LH2 motif in ORF3 of ECF01.GS011.F17 may be involved in mediating membrane attachment via other protein binding partners (Bateman and Sandford, 1999; Tomchick *et al*., 2001) and another motif TonB box, conserved site in ORF5. It is a short conserved region involved in the interaction of the protein with the tonB protein. In *E. coli*, the tonB protein interacts with the outer membrane receptor proteins that were responsible for active transport of specific substrates into the periplasmic space (Gudmundsdottir *et al*., 1989). Thus, this EST may have a function in mediating interaction of membrane proteins.

The alpha/beta hydrolase fold in the ORF2 of ECF02.GS010.K23 and the acid cluster protein 33, can be found in a number of hydrolytic enzymes from different phylogenetic origins and catalytic function (Ollis *et al*., 1992). The acid cluster protein 33 (ACP33) has also been shown to be involved in the intracellular transportation and receptor-mediated signaling pathway (Simpson *et al*., 2003). Another EST was identified with serpin like motifs. Serpins are involved in serine-type endopeptidase inhibitor activity and plays a role in various biological processes such as blood coagulation, fibrinolysis, angiogenesis, inflammation, tumour suppression and hormone transport (Van Gent *et al*., 2003). Based on these motifs, this EST was grouped into cellular organisation.

The class A (penicillinase-type) of motif was present in ECF03.GS001A.A04. This is the most common among the four classes of penicillinase (Knox and Moews, 1991). There were three ESTs (ECM01.GS010.E08 [ORF2], ECF02.GS010.E23 [ORF2] and ECF03.GS002D.B08 [ORF1]) that contained the heavy metal transport/detoxification protein (Bull and Cox, 1994) and copper transport protein

ATOX1-related (Lin and Culotta, 1995). These motifs played a role in resistance against toxic metals and cellular antioxidant defence. Besides being categorised into this functional class, ECM01.GS010.E08 was also grouped into miscellaneous functional class based on the motif detected within ORF6.

ECM01.GS001.F02, ECF01.GS011.E14, ECM01.GS010.J07 and ECM04.GS002D.G12 contain a Mss4-like motif that represents a structural domain which can be found in all Mss4 (Zhu *et al*., 2001). This domain translationally controls tumour-associated protein TCTP (Thaw *et al*., 2001) and the C-terminal MsrB domain of methionine sulphoxide reductase (Lowther *et al*., 2002). Therefore, this EST was assigned a function as a protectant against oxidative damage through its reductase activity (Lowther *et al*., 2000). It was also the same scenario for ORF3 of ECF01.GS011.E14, where universal stress protein UspA and Rossmann-like alpha/beta/alpha sandwich fold were detected. The universal stress protein UspA is a cytoplasmic protein which enhanced the cell survival rate during prolonged exposure to stress agents (Nystrom and Neidhardt, 1994). While, Rossmann-like alpha/beta/alpha sandwich fold,a domain with unknown function, can be found in various protein families, including the universal stress protein family, UspA. Therefore this EST was assigned into a functional class that managed stress endurance.

Motifs sigma factor SIGB regulation protein RSBQ and alpha/beta hydrolases were found on ECM01.GS010.J07 indicating a role in immunity and defence (Brody *et al*., 2001; Kazmierczak *et al*., 2005). Contig 3 and ECM01.GS001.L01 and other ESTs were clustered under the development and structure functional class. Gonadotropins (or glycoprotein hormones) belong to a protein family which includes the mammalian hormones follitropin (FSH), lutropin (LSH), thyrotropin (TSH) and chorionic gonadotropin (CG) (Pierce and Parsons, 1981; Stockell Hartree and Renwick, 1992). The rapid alkalinisation factor (RALF) motif found in ORF2 of ECM01.GS001.L01 is a plant polypeptide that arrest root growth and development (Pearce *et al*., 2001). The other two motifs grouped in this functional class were peroxisome membrane protein, Pex16 and TB2/DP1 and HVA22 related protein.

In this study, we analysed all the clones that were in Category 2 and 4 of the four EST Categories that was generated. The objective of identifying motifs within the ESTs and assigning a putative function was achieved within this research. However, we were not able to determine a putative function for all genes as there are certain motifs that had no assigned function. Though, this library was generated from floral organs, both these categories contained no flowering genes. This was probably due to the fact that only a small number of the clones from both libraries were randomly selected and sequenced for analysis. The percentage of clones in each functional class will be proportionally related to the number of clones sequenced i.e., more clones sequenced the higher the chances of obtaining more useful genes from the library and a better representation of genes in each functional class.

ACKNOWLEDGMENTS

We would like to thank the Ministry of Science Technology and Innovation for the Intensified Research Priority Grant (09-02-02-0036-EA132-Sex determination in *Calamus manan*) that was awarded to conduct this research. This research was also supported by The Malaysian Toray Science Foundation. Both grants were awarded to Dr. Choong Chee Yen.

REFERENCES

Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman, 1990. Basic local alignment search tool. J. Mol. Biol., 215: 403-410.

Andrade, M.A., C. Perez-iratxeta and C.P. Ponting, 2001. Protein repeats: Structures, functions and evolution. J. Struct. Biol., 134: 117-131.

Apweiler, R., T.K. Attwood, A. Bairoch, A. Bateman and E. Birney, 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res., 29: 37-40.

Attwood, T.K., P. Bradley, D.R. Flower, A. Gaulton and N. Maudling, 2003. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res., 31: 400-402.

Bandziulis, R.J., M.S. Swanson and G. Dreyfuss, 1989. RNA-binding proteins as developmental regulators. Genes Dev., 3: 431-437.

Bateman, A. and R. Sandford, 1999. The plat domain: A new piece in the PKD1 puzzle. Curr. Biol., 9: R588-R590.

Bendtsen, J.D., H. Nielsen, G. von Heijne and S. Brunak, 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol., 340: 783-795.

Borden, K.L. and P.S. Freemont, 1996. The RING finger domain: A recent example of a sequence-structure family. Curr. Opin. Struct. Biol., 6: 395-401.

Bouhouche, N., M. Syvanen and C.I. Kado, 2000. The origin of prokaryotic C_2H_2 zinc finger regulators. Trends Microbiol., 8: 77-81.

Brody, M.S., K. Vijay and C.W. Price, 2001. Catalytic function of an a/ß hydrolase is required for energy stress activation of the s^B transcription factor in *Bacillus subtilis*. J. Bacteriol., 183: 6422-6428.

Bru, C., E. Courcelle, S. Carrere, Y. Beausse, S. Dalmar and D. Kahn, 2005. The ProDom database of protein domain families: More emphasis on 3D. Nucleic Acids Res., 33: D212-D215.

Bull, P.C. and D.W. Cox, 1994. Wilson disease and Menkes disease: New handles on heavy-metal transport. Trends Genet., 10: 246-252.

Carr, S., C.N. Penfold, V. Bamford, R. James and A.M. Hemmings, 2000. The structure of TolB, an essential component of the tol-dependent translocation system and its protein-protein interaction with the translocation domain of colicin E9. Structure, 8: 57-66.

Carr, G., N. Simmons and J. Sayer, 2003. A role for CBS domain 2 in trafficking of chloride channel CLC-5. Biochem. Biophys. Res. Commun., 310: 600-605.

Chenuil, A., M. Solignac and M. Bernard, 1997. Evolution of the large-subunit ribosomal RNA binding site for protein L23/25. Mol. Biol. Evol., 14: 578-588.

D'andrea, L.D. and L. Regan, 2003. TPR proteins: The versatile helix. Trends Biochem. Sci., 28: 655-662.

Downing, A.K., V. Knott, J.M. Werner, C.M. Cardy, I.D. Campbell and P.A. Handford, 1996. Solution structure of a pair of calcium-binding epidermal growth factor-like domains: Implications for the Marfan syndrome and other genetic disorders. Cell, 85: 597-605.

Dransfield, J., 2001. Taxonomy, biology and ecology or rattan. Unasylva, 52: 205-205.

Dreyfuss, G., M.S. Swanson and S. Pinol-Roma, 1988. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. Trends Biochem. Sci., 13: 86-91.

Finn, R.D., J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones and V. Hollich, 2006. Pfam: Clans, web tools and services. Nucleic Acids Res., 34: D247-D251.

Freemont, P.S., 1993. The RING finger. A novel protein sequence motif related to the zinc finger. Ann. New York Acad. Sci., 684: 174-192.

Goebl, M. and M. Yanagida, 1991. The TPR snap helix: A novel protein repeat motif from mitosis to transcription. Trends Biochem. Sci., 16: 173-177.

Gough, J., K. Karplus, R. Hughey and C. Chothia, 2001. Assignment of homology to genome sequences using a library of Hidden Markov models that represent all proteins of known structure. J. Mol. Biol., 313: 903-10.1006/jmbi.2001.5080.

Gudmundsdottir, A., P.E. Bell, M.D. Lundrigan, C. Bradbeer and R.J. Kadner, 1989. Point mutations in a conserved region (TonB box) of *Escherichia coli* outer membrane protein BtuB affect vitamin B12 transport. J. Bacteriol., 171: 6526-6533.

Haft, D.H., J.D. Selengut and O. White, 2003. The TIGRFAMs database of protein families. Nucleic Acids Res., 31: 371-373.

Ho, C.L., Y.Y. Kwan, M.C. Choi, S.S. Tee and W.H. Ng, 2007. Analysis and functional annotation of expressed sequence tags (ESTs) from multiple tissues of oil palm (*Elaeis guineensis* Jacq.). BMC Genomics, 8: 381-381.

Hulo, N., A. Bairoch, V. Bulliard, L. Cerutti and E. De Castro, 2006. The PROSITE database. Nucleic Acids Res., 34: D227-D230.

Hunger, S., G. Di Gaspero, S. Möhring, D. Bellin and R. Schäfer-Pregl, 2003. Isolation and linkage analysis of expressed disease-resistance gene analogues of sugar beet (*Beta vulgaris* L.). Genome, 46: 70-78.

Kawaoka, S., S. Katsuma, T. Daimon, R. Isono and N. Omuro, 2008. Functional analysis of four gloverin-like genes in the silkworm, *Bombyx mori*. Arch. Insect. Biochem. Physiol., 67: 87-96.

Kazmierczak, M.J., M. Wiedmann and K.J. Boor, 2005. Alternative sigma factors and their roles in bacterial virulence. Microbiol. Mol. Biol. Rev., 69: 527-543.

Kielkopf, C.L., S. Lucke and M.R. Green, 2004. U2AF homology motifs: Protein recognition in the RRM world. Genes Dev., 18: 1513-1526.

Klug, A., 1999. Zinc finger peptides for the regulation of gene expression. J. Mol. Biol., 293: 215-218.

Knox, J.R. and P.C. Moews, 1991. Beta-lactamase of *Bacillus licheniformis* 749/C. Refinement at 2 A resolution and analysis of hydration. J. Mol. Biol., 220: 435-455.

Krogh, A., B. Larsson, G. Von Heijne and E.L.L. Sonnhammer, 2001. Predicting trans-membrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol., 305: 567-580.

Laity, J.H., B.M. Lee and P.E. Wright, 2001. Zinc finger proteins: New insights into structural and functional diversity. Curr. Opin. Struct. Biol., 11: 39-46.

Lazo, G.R., S. Chao, D.D. Hummel, H. Edwards and C.C. Crossman, 2004. Development of an expressed sequence tag (EST) resource for wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map. Genetics, 168: 585-593.

Letunic, I., R.R. Copley, B. Pils, S. Pinkert, J. Schultz and P. Bork, 2006. SMART 5: Domains in the context of genomes and networks. Nucleic Acids Res., 34: D257-D260.

Li, D. and R. Roberts, 2001. WD-repeat proteins: Structure characteristics, biological function, and their involvement in human diseases. Cell Mol. Life Sci., 58: 2085-2097.

Lin, S.J. and V.C. Culotta, 1995. The ATX1 gene of *Saccharomyces cerevisiae* encodes a small metal homeostasis factor that protects cells against reactive oxygen toxicity. Proc. Natl. Acad. Sci. USA., 92: 3784-3788.

Lowther, W.T., N. Brot, H. Weissbach, J.F. Honek and B.W. Matthews, 2000. Thiol-disulfide exchange is involved in the catalytic mechanism of peptide methionine sulfoxide reductase. Proc. Natl. Acad. Sci. USA., 97: 6463-6468.

Lowther, W.T., H. Weissbach, F. Etienne, N. Brot and B.W. Matthews, 2002. The mirrored methionine sulfoxide reductases of *Neisseria gonorrhoeae* pilB. Nat. Struct. Biol., 9: 348-352.

Matthews, J.M. and M. Sunde, 2002. Zinc fingers-folds for many occasions. IUBMB Life, 54: 351-355.

Mi, H., B.L. Ulitsky, R. Loo, A. Kejariwal and J. Vandergriff *et al*., 2005. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res., 33: D284-D288.

Mueckler, M., C. Caruso, S.A. Baldwin, M. Panico and I. Blench *et al*., 1985. Sequence and structure of a human glucose transporter. Science, 229: 941-945.

Munro, S. and H.R.B. Pelham, 1987. A C-terminal signal prevents secretion of luminal ER proteins. Cell, 48: 899-907.

Nadarajah, K., C.Y. Choong, W. Ratnam, S.J. Leong, B.K. Thi, P. Hedley and R. Waugh, 2006. EST analysis in *Calamus manan* Miq. J. Trop. Plant Physiol., 1: 1-11.

Nadarajah, K., L.S. Jye, C.C. Yen and W. Ratnam, 2009. Identification of floral ESTs from *Calamus manan* inflorescence library. Biosci Biotechnol. Res. Asia.

Nystrom, T. and F.C. Neidhardt, 1994. Expression and role of the universal stress protein, UspA, of *Escherichia coli* during growth arrest. Mol. Microbiol., 11: 537-544.

Ollis, D.L., E. Cheah, M. Cygler, B. Dijkstra and F. Frolow *et al*., 1992. The alpha/beta hydrolase fold. Protein Eng., 5: 197-211.

Pearce, G., D.S. Moura, J. Stratmann and C.A. Ryan Jr., 2001. RALF, a 5-kDa ubiquitous polypeptide in plants, arrests root growth and development. Proc. Natl. Acad. Sci. USA., 98: 12843-12847.

Pelham, H.R.B., 1990. The retention signal for soluble proteins of the endoplasmic reticulum. Trends Biochem. Sci., 15: 483-486.

Pierce, J.G. and T.F. Parsons, 1981. Glycoprotein hormones: Structure and function. Annu. Rev. Biochem., 50: 465-495.

Prather, D., N.J. Krogan, A. Emili, J.F. Greenblatt and F. Winston, 2005. Identification and characterisation of *Elf1*, a conserved transcription elongation factor in *Saccharomyces cerevisiae*. Mol. Cell. Biol., 254: 10122-10135.

Preker, P.J. and W. Keller, 1998. The HAT helix, a repetitive motif implicated in RNA processing. Trends Biochem. Sci., 23: 15-16.

Query, C.C., R.C. Bentley and J.D. Keene, 1989. A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. Cell, 57: 89-101.

Rost, B. and A. Valencia, 1996. Pitfalls of protein sequence analysis. Curr. Opin. Biotechnol., 7: 457-461.

Sachs, A.B., R.W. Davis and R.D. Kornberg, 1987. A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability. Mol. Cell. Biol., 7: 3268-3276.

Sciara, G., S.G. Kendrew, A.E. Miele, N.G. Marsh and L. Federici *et al*., 2003. The structure of ActVA-Orf6, a novel type of monooxygenase involved in actinorhodin biosynthesis. EMBO J., 22: 205-215.

Scott, J.W., S.A. Hawley, K.A. Green, M. Anis and G. Stewart *et al*., 2004. CBS domains form energy-sensing modules whose binding of adenosine ligands is disrupted by disease mutations. J. Clin. Invest., 113: 274-284.

Simpson, M.A., H. Cross, C. Proukakis, A. Pryde and R. Hershberger *et al*., 2003. Maspardin is mutated in mast syndrome, a complicated form of hereditary spastic paraplegia associated with dementia. Am. J. Hum. Genet., 73: 1147-1156.

Stockell Hartree, A. and A.G. Renwick, 1992. Molecular structures of glycoprotein hormones and functions of their carbohydrate components. Biochem. J., 287: 665-679.

Tan, J., H.L. Wang and K.W. Yeh, 2005. Analysis of organ-specific, expressed genes in *Oncidium* orchid by subtractive expressed sequence tags library. Biotechnol. Lett., 27: 1517-1528.

Thaw, P., N.J. Baxter, A.M. Hounslow, C. Price, J.P. Waltho and C.J. Craven, 2001. Structure of TCTP reveals unexpected relationship with guanine nucleotide-free chaperones. Nat. Struct. Biol., 8: 701-704.

Tomchick, D.R., P. Phan, M. Cymborowski, W. Minor and T.R. Holman, 2001. Structural and functional characterisation of second-coordination sphere mutants of soybean lipoxygenase-1. Biochemistry, 40: 7509-7517.

Torto-Alalibo, T.A., S. Tripathy, B.M. Smith, F.D. Arredondo and L. Zhou *et al*., 2007. Expressed sequence tags from phytophthora sojae reveal genes specific to development and infection. Mol. Plant Microbe Interact., 20: 781-793.

Van Gent, D., P. Sharp, K. Morgan and N. Kalsheker, 2003. Serpins: Structure, function and molecular evolution. Int. J. Biochem. Cell Biol. 35: 1536-1547.

Wang, L., Y. Xu, C. Zhang, Q. Ma and S.H. Joo *et al*., 2008. OsLIC, a novel CCCH-type zinc finger protein with transcription activation, mediates rice architecture via brassinosteroids signaling. PLoS ONE, 3: e3521-e3521.

Wise, C.A., L.C. Chiang, W.A. Paznekas, M. Sharma and M.M. Musy *et al*., 1997. TCOF1 gene encodes a putative nucleolar phosphoprotein that exhibits mutations in Treacher Collins syndrome throughout its coding region. Proc. Natl. Acad. Sci. USA., 94: 3110-3115.

Wu, C.H., A. Nikolskaya, H. Huang, L.S. Yeh and D.A. Natale *et al*., 2004. PIRSF: Family classification system at the protein information resource. Nucleic Acids Res., 32: D112-D114.

Yeats, C., M. Maibaum, R. Marsden, M. Dibley, D. Lee, S. Addou and C.A. Orengo, 2006. Gene3D: Modelling protein structure, function and evolution. Nucleic Acids Res., 34: D281-D284.

Zdobnov, E.M. and R. Apweiler, 2001. InterProScan-an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 17: 847-848.

Zhu, Z., J.J. Dumas, S.E. Lietzke and D.G. Lambright, 2001. A helical turn motif in Mss4 is a critical determinant of *Rab* binding and nucleotide release. Biochemistry, 40: 3027-3036.